DOCUMENT RESUME

ED 095 216 TM 003 889

AUTHOR Hogan, Thomas P.; Beck, Michael
TITLE Comparability of Different Score Modes for Analysis of Group Data.
PUB DATE [Apr 74]
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Comparative Analysis; Correlation; Grade 4; Groups; Raw Scores; *Scores; *Standardized Tests; Statistical Data; Test Interpretation; Test Reliability; *Test Results; Tests of Significance
IDENTIFIERS Metropolitan Achievement Tests

ABSTRACT

Scores on standardized tests may be expressed in a variety of different modes. Raw scores, grade equivalents, percentile ranks, stanines, and standard scores are the major types in use today. The study investigated whether the mode or scale used to express scores affected the magnitude of group statistics, specifically, t and r (the t-test of the significance of differences between means and inter-test correlations). Raw score data were generated randomly, then translated into grade equivalents, percentile ranks, stanines, and standard scores using norm tables from Metropolitan Achievement Tests. This procedure was replicated five times for three levels of t and r, for three sample sizes, for two tests. Conclusions were drawn regarding when the use of a converted score yielded t or r values that were noticeably different from t or r values obtained when using raw scores. (Author)

(American Educational Research Association, Annual Meeting, Chicago, 1974 Session 24.02)

ED 095216

U S DEPARTMENT OF HEALTH EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Comparability of Different Score Modes for Analysis of Group Data

BEST COPY AVAILABLE

Thomas P. Hogan
University of Wisconsin-Green Bay

BEST COPY AVAILABLE

Michael Beck
Fordham University

PROBLEM

Much educational research is carried out with standardized tests, the results of which may be expressed in terms of several different score modes, e.g. raw scores, grade equivalents, percentile ranks, stanines, or standard scores.

The respective advantages and disadvantages of these score scales in the interpretation of individual performance are relatively well-known. Descriptions of the characteristics of various scales as they relate to individual performance can be found in standard textbooks on testing, e.g. Anastasi (1968), Cronbach (1970), or Lyman (1971).

In many educational research projects, the concern is not with the interpretation of individual performance but with group results. In this circumstance, the question is frequently raised as to whether the results of analysis of group performance depend in any way on the score scale used. For example, is one any more or less likely to identify a significant difference in the achievement of two groups if results are expressed in terms of percentile ranks rather than in raw scores? Or, if a test manual states that the correlation between raw scores on two tests is .60, can it be assumed that the correlation would be .60 if scores were expressed in terms of stanines?

The purpose of the present study was to investigate differences in group statistics as a consequence of the type of scale used to express the scores. The score modes compared were raw score (RS), grade equivalent (GE), percentile rank (PR),

TM 003 389

BEST COPY AVAILABLE

stanine (S), and Thurstone-type standard score (SS). These scales are the ones used primarily for expressing the results of standardized educational tests.

The question of whether the score scale used in statistical analyses affects the results may be very critical for the researcher in a number of different instances. Perhaps most importantly, data may be available to the researcher in only one form, e.g. in grade equivalents, and it may be impossible or extremely difficult to recover the original raw scores. Or the researcher may have access to records which give a number of different transformed scores and he must choose the type of score scale to be used for analysis. Of a somewhat less critical nature, but especially important in large scale research projects, the researcher may wish to choose the score scale to be used in terms of simplicity of data processing. For example, stanines are obviously much easier to work with than are other score scales in terms of the sheer volume of digits recorded. In each of these instances, some estimate of the effect of using one score scale rather than another would be useful.

The general question about whether group statistics are affected by the type of score scale used was resolved into three more specific questions:

1) When correlating one set of test scores expressed in terms of any score mode and another set of scores expressed in a fixed scale, will the correlation coefficient be affected by the score mode used for the first test? For example, will the correlation between raw scores on Test A and IQ be approximately the same as the correlation between stanines on Test A and IQ?

2) When correlating two sets of test scores, both of which are expressed in the same score mode, will the correlation coefficient be about the same regardless of the mode used to express the two sets of scores? For example, will the correlation between Reading scores obtained at the beginning and end of some experimental program be about the same whether the two sets of scores are expressed as grade equivalents or as raw scores?

3) When testing for significance of the difference between two means, does the score scale used affect the level of t? For example, is one less likely to detect a significant difference if scores are expressed as stanines rather than as raw scores?

METHOD

Each of the above three questions was approached using three different sample sizes and three different "levels" of the statistic (t or r) of interest. For all analyses, sample sizes of 25, 100, and 1000 were used. For questions 1 and 2, r levels of approximately .50, .75, and .90 were used. These values seem to represent the range of inter-test correlations ordinarily found when working with the types of tests involved in the present study. For analysing question 3, the three approximate levels of t used were 1.00, i.e. a small, clearly insignificant difference; 2.00, i.e. a moderate or borderline-significant level of difference; and 5.00, i.e. a substantial, clearly significant difference for the sample sizes used here.

Score distributions for selected tests of the Metropolitan Achievement Tests, 1970 Edition, were used as the parameters for generating artificial raw score distributions. This test series was selected since it includes all of the above-mentioned types of scores in its regular manuals and reporting services. (For a description of the development of norms and score scales for this test series, see Metropolitan Achievement Tests, 1970a, 1970b.)

The specific tests within the Metropolitan series which were used in the present study were the Total Reading test and the Mathematics Computation test of the Elementary battery. Norm distributions for the beginning of Grade 4 were used for the percentile ranks and stanines needed for the present study. The Total Reading test is composed of a 50 item Work Knowledge subtest and a 45 item

Reading Comprehension subtest. The Computation test contains 40 items. Split-half reliabilities for the Total Reading and Math Computation tests at the beginning of Grade 4 are .96 and .88, respectively (Metropolitan Achievement Tests, 1970b). These two tests were selected for analyses since they differ noticeably in length, reliability, and curricular area covered. Grade 4 was chosen simply because it represents a grade near the middle of the recommended grade levels for the Metropolitan series.

The univariate and bivariate distribution of scores required for analysis and the appropriate statistics were generated via computer. Fortran programs written expressly for this purpose utilized the IBM Scientific Subroutine Package subroutine Gauss, which generates random normal numbers. All computations were done on the Fordham University IBM 360-40 computer.[1] Desired raw score means and standard deviations and either (a) levels of difference between groups (for the t analyses) or (b) approximate levels of correlation between tests (for the r analyses) were specified for the raw score data. Then distributions of scores meeting these requirements were randomly generated, the resulting raw scores were translated into grade equivalents, standard scores, percentile ranks, and stanines using the published Metropolitan norm tables, which had been read into the computer, and the desired statistics (t or r) were calculated.

The above procedure was carried out independently for each level of N (25, 100, 1000) for both the Total Reading and Computation tests. In addition, the entire procedure was independently replicated five times for each level of N for each test. Thus, a total of 90 separate determinations were made in analysing each of the questions posed above: 3 levels of N x 3 levels of t (or r) x 2 tests x 5 replications.

All data required for analyses relating to the second and third questions listed above were derived from the Metropolitan norm tables. For the first question one "dummy" variable was needed. The "dummy" variable was arbitrarily assigned a

mean of 100 and standard deviation of 15; the actual values used for the analyses were again randomly generated. For convenience, this variable may be thought of as IQ, although it might be family income, rate of reading, or any other variable for which the scale remains constant while it is being correlated with another variable expressed in RS, GE, PR, S, or SS.

RESULTS

The primary concern in this study was with differences in the t or r values obtained when using the different score scales. Of the various score scales, the raw score scale is the most basic since all other scales are transformations of the raw score scale. Therefore, results are presented in terms of deviations of the t or r values obtained when using raw scores. In evaluating the magnitude of these deviations, it is not clear whether one should consider the average deviation, taking into account the direction of the deviation, or the absolute average deviation as the most valid measure of discrepancy. Therefore, both the simple average deviation and the absolute average deviation are reported for all analyses. Averages are taken across the five replications within each type of analysis. Absolute average deviations are represented as $/\bar{d}/$ and simple average deviations as $\bar{d}$.

Table 1 presents differences between correlation coefficients based on use of raw scores and based on use of GE, PR, S, and SS for the three levels of r and the three levels of N for the Reading test. Table 2 presents the same type of data for the Math Computation test. The correlation coefficients used to obtain the data in Tables 1 and 2 were based on one test which was expressed successively in each of the different score modes and a "dummy" variable for which the scale did not change.

(Insert Tables 1 and 2)

Tables 3 and 4 present data comparable to that in Tables 1 and 2, respectively, for the case in which scales change for both variables involved in the correlation.

(Insert Tables 3 and 4)

Tables 5 and 6 present differences between t values, to the nearest one-tenth, based on raw scores and the other score modes. Data are presented for three levels of t.

(Insert Tables 5 and 6)

DISCUSSION AND CONCLUSIONS

Examination of data in Tables 1 and 2 suggests the following conclusions regarding the correlation between a test variable expressed in any one of the five score modes and a variable expressed in some fixed score scale:

a) For all score transformations and all levels of N used here, there appears to be no relationship between level of r and magnitude of $\bar{d}$ or $/\bar{d}/$.

b) There is a very slight tendency for $/\bar{d}/$ to be greater for N = 25 than for N = 100 or N = 1000. The $/\bar{d}/$ values for N = 100 and N = 1000 are virtually identical. There appears to be no relationship between $\bar{d}$ and level of N. These findings are reasonable in light of the differential degrees of sampling variability associated with the three levels of N used in the study.

c) When the test variable is expressed in grade equivalents, the correlation is about .05 less than when the test variable is expressed in raw scores. In the Reading data there is a relationship between the degree of discrepancy and level of r but this relationship is not confirmed in the Math Computation data.

d) When the test variable is expressed in Stanines the correlation is about .01 or .02 less than when the test variable is expressed in raw scores.

e) When the test variable is expressed in Percentile Ranks the correlation is about .01 less than when the test variable is expressed in raw scores.

f) It is difficult to arrive at a single conclusion for the Standard Score data, since the Reading and Math Computation values differ considerably. Considering only the Computation data, it appears that expressing the test variable in Standard Scores yields r's which differ negligibly from r's obtained when the test variable is expressed in raw scores. Considering only the Reading test data, it appears that expressing the test variable in Standard Scores yields r's which are on the average about .05 lower than r's obtained when the test variable is expressed in raw scores.

Examination of data in Tables 3 and 4 suggests the following conclusions regarding the correlation between two test variables both of which are expressed in the same score mode:

a) For all score transformations and all levels of N, there appears to be no relationship between levels of r and $\bar{d}$ or $/\bar{d}/$. This is consistent with the conclusion reached above for data in Tables 1 and 2.

b) For all score transformations and all levels of r, there is a slight tendency for $/\bar{d}/$ to be greater for N = 25 than for N = 100 and N = 1000. The $/\bar{d}/$ values for N = 100 and N = 1000 are, on the average, almost identical. There appears to be no relationship between $\bar{d}$ and N. Again, as in conclusion (b) above, this finding was as expected due to sampling.

c) If both test variables are expressed in PR, the correlation will be about .01 or .02 less than if both variables are expressed in raw scores.

d) For the comparison of GE, S, and SS with RS, the data are inconsistent for the Reading and Computation tests. For Computation, the differences tend to be

small - about .02. For Reading, the differences are usually at least .05 and as high as .13. It must be remembered that the data referred to here are averages across five replications; actual deviations for some samples were even larger.

Examination of data in Tables 5 and 6 suggests the following conclusions regarding the t values obtained when the test variables are expressed as GE, PR, S, or SS as opposed to raw scores:

a) For all four transformations there are many instances in which the obtained t value is actually greater than the t value obtained when raw scores are used. However, $\bar{d}$ is negative in the majority of instances; the overall tendency is towards lower t values when using transformed scores as opposed to raw scores.

b) There appears to be no relationship between sample size and $\bar{d}$ or $/\bar{d}/$ for any of the score transformations.

c) There appears to be no relationship between the level of t and $\bar{d}$ or $/\bar{d}/$. While there is a suggestion of such a relationship (positive) for the Computation data, the fact that the relationship is not strong and is not confirmed in the Reading data suggests that the trend in the Computation data is of no practical consequence.

d) The t values obtained when the PR transformation is used are very close approximations to those obtained when raw scores are used. Considering PR data for all levels of t and N in Tables 5 and 6, one would expect a t based on PR scores to differ from a t based on raw scores by about -.1. Such a difference generally has little practical significance.

e) The t values obtained when S or SS transformations are used tend to deviate from t values obtained when using raw scores by about -.2. Both the $\bar{d}$ and $/\bar{d}/$ values for S and SS are rather consistently larger than for PR. However, as for the PR data, the differences in t values for S and SS as opposed to t values for raw scores seem to have little practical significance.

f) The *t* values obtained when the GE transformation is used deviate by as much as -.9, on the average, from *t* values obtained when using raw scores. Considering all of the data for the GE transformation in Tables 5 and 6, it would appear that use of GE scores instead of raw scores when doing a *t* test may have a noticeable effect on the magnitude of *t*. Specifically, if GE scores are used in the analysis, an investigator may conclude that he has not obtained a significant difference when in fact a significant difference might have been obtained had he used raw scores. (In statistical terms, this means that the use of GEs for *t* test analyses increases the probability of a β error and decreases the statistical power of the *t* test.) Of course, many of the $\bar{d}$ and $/\bar{d}/$ values for the GE data indicate that use of GE scores for computing *t* values will often give close approximations to the *t* values which would be obtained with raw scores. However, the magnitude of some of the $\bar{d}$ and $/\bar{d}/$ values indicates that caution should be exercised in the use of GE scores.

CONCLUSION

In general, considering all of the data presented in this study, the following recommendations regarding the use of the various score modes in place of raw scores seem warranted. Use of percentile ranks rather than raw scores seems to have no important effect on the level of *t* or *r* obtained. Use of stanines rather than raw scores appears to have little practical consequence except when obtaining correlations between two variables, both of which are expressed in stanines; in this instance, the researcher can expect the correlation to be about .05 lower than if both variables had been expressed in raw scores. Use of standard scores rather than raw scores seems to make little difference when computing *t*'s; however, for computing correlations when either one or both variables are expressed in standard scores, the researcher may expect the correlations to be about .05 less than if raw scores are

used. (Even the ".05" value is not a very good guideline since data for standard scores were quite variable.) Use of grade equivalents rather than raw scores requires considerable caution. The t values obtained when using the GE scale are often sufficiently lower than those obtained when using raw scores that conclusions may be affected. The correlations obtained when using grade equivalents for one or both variables may be expected to be about .05 less than if raw scores had been used.

The conclusions reached in the present study should be re-examined in a number of different ways. First, the study should be replicated using other tests. Second, the conclusions should be checked using distributions with various shapes; the distributions (of raw scores) used in the present study were designed to be normal. Third, the conclusions should be checked using restricted portions of the various score scales. For example, would the conclusions still hold if data were for extremely below-average cases? This question would be especially important for the percentile rank and stanine scales. Fourth, the conclusions relating to standard scores should be investigated for other types of standard score transformations; Thurstone-type standard scores were the only type involved in the present study.

References

Anastasi, A. Psychological Testing, 3rd Edition. New York: Macmillan, 1968.

Cronbach, L. J. Essentials of Psychological Testing, 3rd Edition. New York: Harper & Row, 1970.

Lyman, H. B. Test Scores and What They Mean, 2nd Edition. Englewood Cliffs, N. J.: Prentice Hall, 1971.

Metropolitan Achievement Tests. Special Report No. 7: Guidelines for Standardization Sampling. New York: Harcourt Brace Jovanovich, 1970(a).

---- Teacher's Handbook, Elementary. New York: Harcourt Brace Jovanovich, 1970(b).

1. The authors gratefully acknowledge the cooperation of the Fordham University Psychology Department and Dr. John Walsh for granting computer time for this project.

Table 1

Difference Between r's for Three Levels of r and Three Levels of N Based on Use of Raw Scores and Use of GE, S, PR, and SS for Total Reading and a "Dummy" Variable[a]

		r=.50				r=.75				r=.90			
N	Scale:	GE	S	PR	SS	GE	S	PR	SS	GE	S	PR	SS
25	$\bar{d}$	-06	-02	-01	-07	-01	00	-01	-01	-07	-04	-01	-07
	$/\bar{d}/$	07	04	02	07	03	02	01	03	07	04	02	07
100	$\bar{d}$	-02	-01	00	-03	-05	-02	-01	-05	-07	-03	-01	-06
	$/\bar{d}/$	03	01	01	03	05	02	01	05	07	03	01	06
1000	$\bar{d}$	-04	-01	-01	-03	-06	-02	-01	-05	-07	-02	-01	-07
	$/\bar{d}/$	04	01	01	03	06	02	01	05	07	02	01	07

[a]Decimal points omitted.

Table 2

Difference Between r's for Three Levels of r and Three Levels of N Based on Use of Raw Scores and Use of GE, S, PR, and SS for Computation and a "Dummy" Variable[a]

		r=.50				r=.75				r=.90			
N	Scale:	GE	S	PR	SS	GE	S	PR	SS	GE	S	PR	SS
25	$\bar{d}$	-05	00	00	-01	-05	01	00	-01	-03	00	-02	-01
	$/\bar{d}/$	05	01	03	01	06	01	02	02	03	01	02	01
100	$\bar{d}$	-03	00	-02	00	-06	-01	-01	-01	-04	-01	-01	00
	$/\bar{d}/$	03	00	02	01	06	02	01	01	04	01	01	01
1000	$\bar{d}$	-03	-01	-01	00	-04	-01	-01	00	-03	-01	-02	-01
	$/\bar{d}/$	03	01	01	00	04	01	01	01	05	01	02	01

[a]Decimal points omitted.

Table 3

Difference Between r's for Three Levels of r and Three Levels of N Based on Use of Raw Scores and Use of GE, S, PR, and SS for Two Sets of Reading Scores[a]

		r=.50				r=.75				r=.90			
N	Scale:	GE	S	PR	SS	GE	S	PR	SS	GE	S	PR	SS
25	$\bar{d}$	-07	-11	-03	-05	-02	-07	-02	-08	-01	-10	-01	-02
	$/\bar{d}/$	09	13	04	06	09	07	04	11	02	10	01	03
100	$\bar{d}$	-03	-03	-01	-03	-05	-07	-01	-06	-04	-08	-02	-06
	$/\bar{d}/$	03	05	01	04	05	07	01	06	04	08	02	06
1000	$\bar{d}$	-06	-04	-01	-06	-05	-06	-01	-07	-05	-08	-02	-08
	$/\bar{d}/$	06	04	01	06	05	06	01	07	05	08	02	08

[a]Decimal points omitted.

Table 4

Difference Between r's for Three Levels of r and Three Levels of N Based on Use of Raw Scores and Use of GE, S, PR, and SS for Two Sets of Computation Scores[a]

		r=.50				r=.75				r=.90			
N	Scales:	GE	S	PR	SS	GE	S	PR	SS	GE	S	PR	SS
25	$\bar{d}$	-05	-03	00	-03	-01	-01	-04	-04	00	-03	-02	-01
	$/\bar{d}/$	05	05	05	05	02	03	04	04	02	03	02	02
100	$\bar{d}$	-02	-02	-02	00	-03	-03	-01	-03	-02	-03	-02	-02
	$/\bar{d}/$	04	02	03	03	03	03	02	03	02	03	02	02
1000	$\bar{d}$	-01	-01	-02	-01	-02	-02	-02	-01	-02	-02	-01	-01
	$/\bar{d}/$	02	01	02	01	02	02	02	01	02	02	01	01

[a]Decimal points omitted.

Table 5
Difference Between t's (to Nearest One-tenth) for Three Levels of t and Three Levels of N Based on Use of Raw Scores and Use of GE,S, PR, and SS for Reading Scores[a]

		t=1.0				t=2.0				t=5.0			
N	Scales:	GE	S	PR	SS	GE	S	PR	SS	GE	S	PR	SS
25	$\bar{d}$	-4	-2	0	-4	-1	-1	1	0	-2	-2	1	-1
	$/\bar{d}/$	6	3	2	4	2	2	1	3	2	2	1	2
100	$\bar{d}$	-2	-1	-1	-1	1	0	-1	3	-6	-4	-1	-4
	$/\bar{d}/$	3	2	1	1	1	1	1	3	6	4	2	4
1000	$\bar{d}$	-3	-1	-2	-3	-3	-2	0	-3	-3	-1	0	-1
	$/\bar{d}/$	5	3	2	6	4	2	1	4	3	1	1	3

[a]Decimal points omitted.

Table 6
Difference Between t's (to Nearest One-tenth) for Three Levels of t and Three Levels of N Based on Use of Raw Scores and Use of GE, S, PR, and SS for Computation Scores[a]

		t=1.0				t=2.0				t=5.0			
N	Scales:	GE	S	PR	SS	GE	S	PR	SS	GE	S	PR	SS
25	$\bar{d}$	0	-2	0	0	-3	-1	-1	0	-4	-1	-1	0
	$/\bar{d}/$	2	2	1	1	3	2	2	1	5	2	2	1
100	$\bar{d}$	-1	0	-1	0	-2	-1	-1	-1	-7	-1	0	-3
	$/\bar{d}/$	2	1	1	1	2	1	2	2	7	1	1	3
1000	$\bar{d}$	0	-1	-2	1	-2	-1	-1	-1	-9	-2	-2	-2
	$/\bar{d}/$	2	1	2	1	2	2	2	1	9	2	2	2

[a]Decimal points omitted.

THE END